(12) UK Patent Application (19) GB (11) 2 343 348 (13) A

(43) Date of A Publication 03.05.2000

(21) Application No 9828598.4

(22) Date of Filing 23.12.1998

(30) Priority Data
(31) 98041142 (32) 30.09.1998 (33) KR

(51) INT CL[7]
H04N 7/173

(52) UK CL (Edition R)
H4R RCSS

(56) Documents Cited
EP 0770964 A1 EP 0767585 A2 US 5737747 A
US 5630007 A

(58) Field of Search
UK CL (Edition Q) H4K KOD3 , H4R RCSS RCST RCT RCX
INT CL[6] H04N 7/173
ONLINE - EPODOC, WPI

(71) Applicant(s)
Daewoo Electronics Co., Ltd.
(Incorporated in the Republic of Korea)
541 5-Ga, Namdaemoon-Ro, Jung-Ku, Seoul,
Republic of Korea

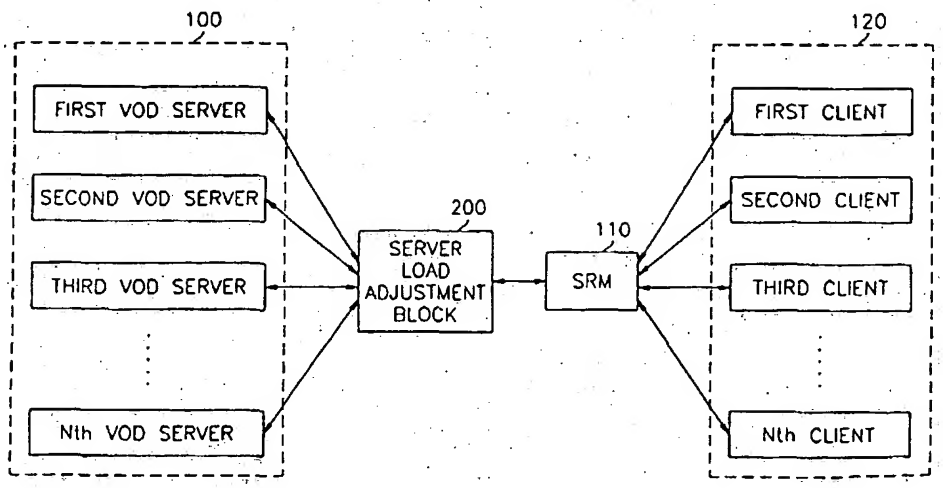
(72) Inventor(s)
Tae-Beom Lim

(74) Agent and/or Address for Service
Page White & Farrer
54 Doughty Street, LONDON, WC1N 2LS,
United Kingdom

(54) Abstract Title
Server load balancing in a video on demand system

(57) A video on demand (VOD) system has a plurality of VOD servers 100, a multiplicity of clients 120, a Session and Resource Manager 110, and a server load adjustment/balancing block 200. The block 200 has a server status information block (350, Fig.3) which stores a maximum transmission rate and a current transmission rate for each server 100. When a service request message is received from a client 120, the load balancing block 200 determines the remnant transmission rate for each server in turn, the remnant transmission rate being calculated by subtracting the current transmission rate from the maximum transmission rate of the server under consideration. When a server is found which has a remnant transmission rate exceeding the transmission rate required to provide a video stream corresponding to the service request message, the service request message is assigned to that server and a session-setup-response message is transmitted to the Session and Resource Manager 110.

FIG. 2



GB 2 343 348 A

FIG. 1
(PRIOR ART)

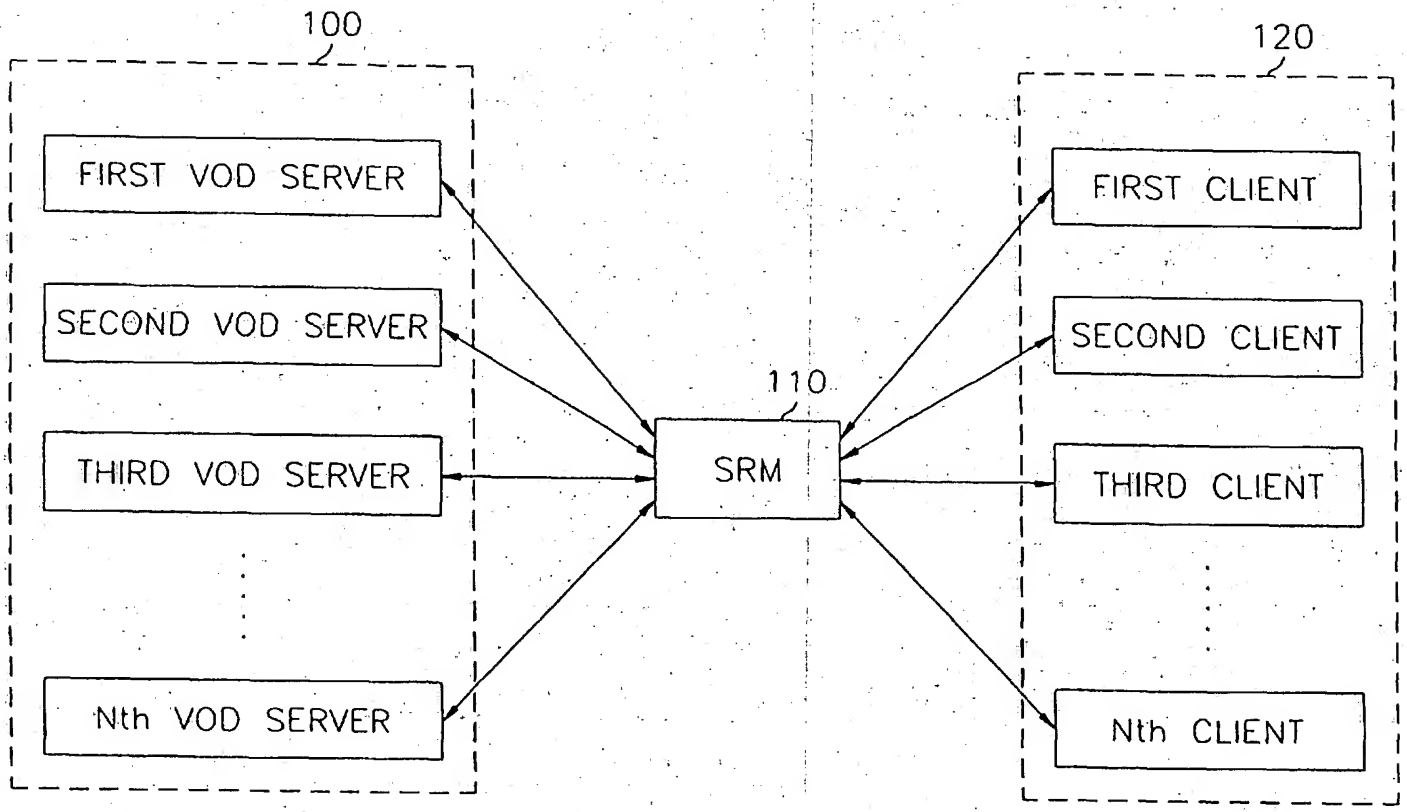


FIG. 2

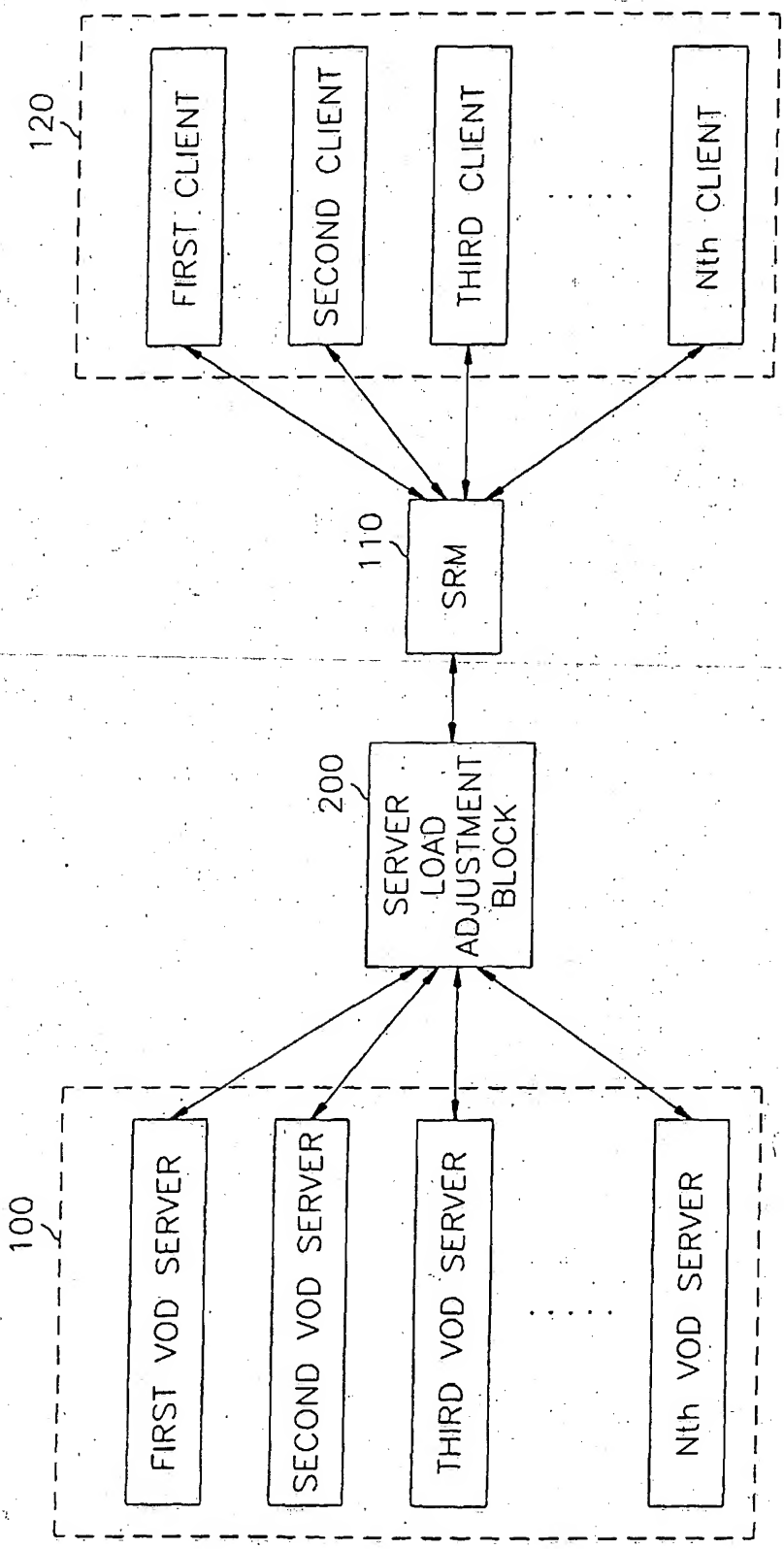

120
FIRST CLIENT
SECOND CLIENT
THIRD CLIENT
Nth CLIENT
110
SRM
200
SERVER LOAD ADJUSTMENT BLOCK
100
FIRST VOD SERVER
SECOND VOD SERVER
THIRD VOD SERVER
Nth VOD SERVER

FIG. 3

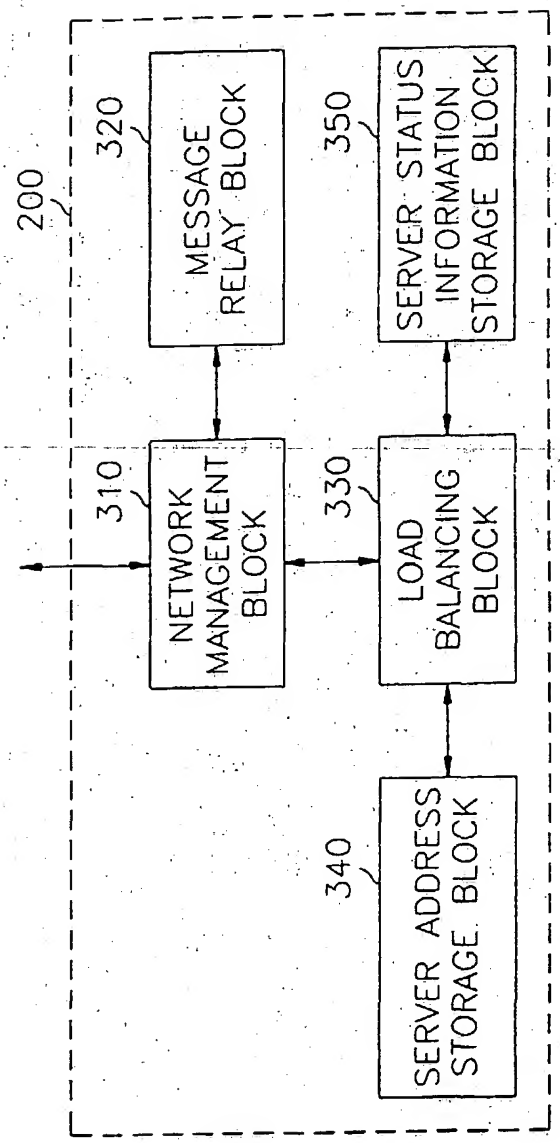
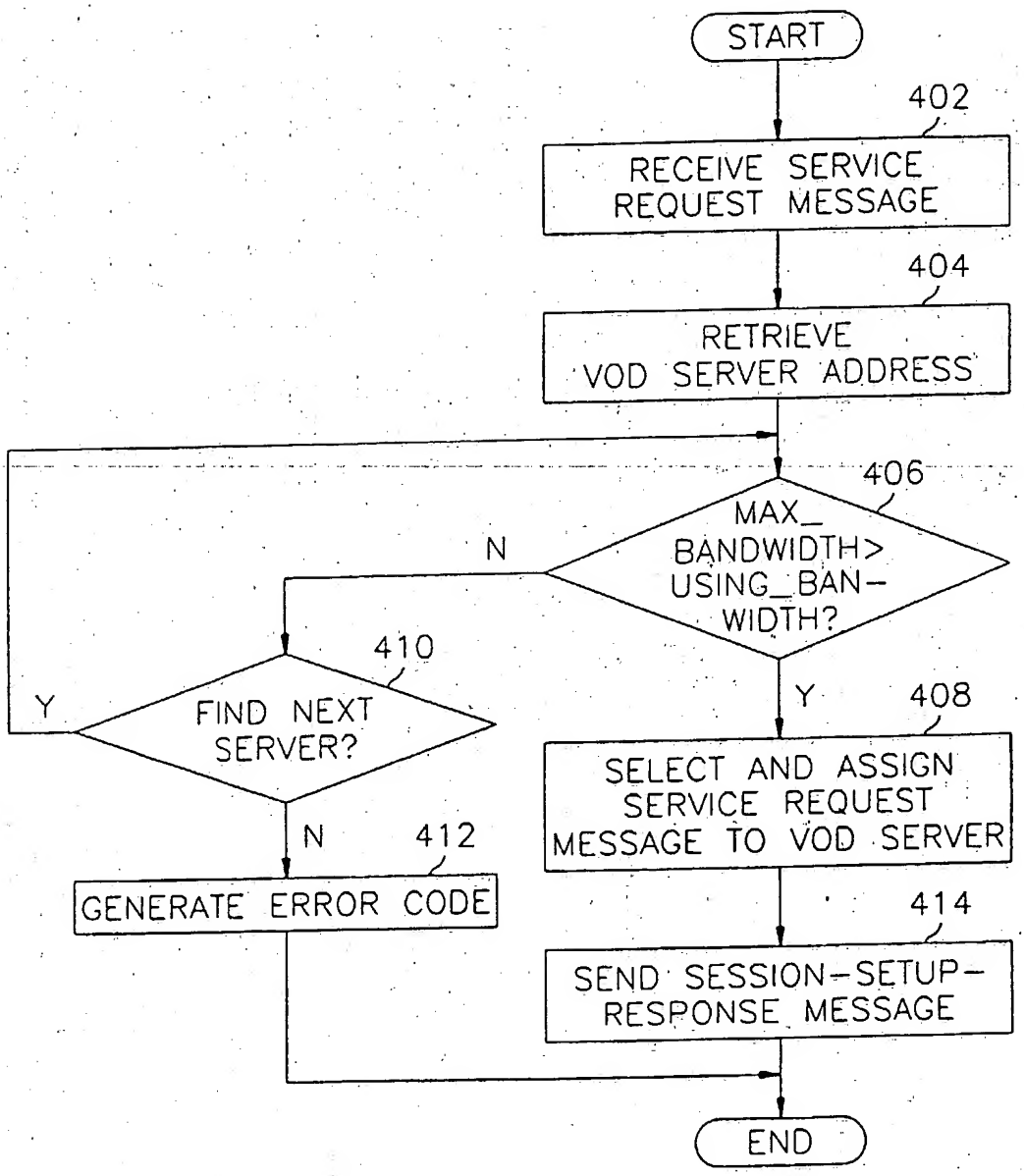

200
310
NETWORK MANAGEMENT BLOCK
320
MESSAGE RELAY BLOCK
330
LOAD BALANCING BLOCK
340
SERVER ADDRESS STORAGE BLOCK
350
SERVER STATUS INFORMATION STORAGE BLOCK

FIG. 4


START
402
RECEIVE SERVICE REQUEST MESSAGE
404
RETRIEVE VOD SERVER ADDRESS
406
MAX_BANDWIDTH> USING_BAN-WIDTH?
N
Y
408
SELECT AND ASSIGN SERVICE REQUEST MESSAGE TO VOD SERVER
410
FIND NEXT SERVER?
Y
N
412
GENERATE ERROR CODE
414
SEND SESSION-SETUP-RESPONSE MESSAGE
END

METHOD AND APPARATUS FOR BALANCING A LOAD OF A SERVER IN A VIDEO ON DEMAND SYSTEM

The present invention relates to a video on demand (VOD) system; and more particularly, to a method and apparatus for balancing a load of a server in a VOD system.

A conventional VOD system is usually equipped with a video storage block storing a plurality of video programs in the form of a video stream, a service manager and a network interface. In the VOD system, the video programs, responsive to service request messages from a plurality of clients, are retrieved from the video storage block under the control of the service manager; and then the retrieved video programs are supplied to the clients through the network interface.

In order to standardize the conventional VOD system, ISO (International Organization for Standardization) prescribes a service manager, e.g., a digital storage media command and control (DSM-CC). The DSM-CC is a set of protocols prescribing a common management and control function in order to manage data, e.g., a plurality of video programs, stored in the video storage block regardless of the type of the digital storage medium. Such DSM-CC exists in a server-to-client environment transferring data between a client and a

server and in a user-to-network environment transferring data between a network and a client or a network and a server.

Referring to Fig. 1, there is shown a conventional distributed VOD system comprising a plurality of VOD servers 100, a SRM (Session and Resource Manager) 110 and a multiplicity of clients 120. Each of the plurality of servers 100 is implemented with a low cost equipment. The SRM 110 connects one of the clients 120 requesting a service to one of the VOD servers 100 capable of providing the requested service by examining a load status of the VOD server.

Since, however, the conventional distributed VOD system has shortcomings in that each of the plurality of VOD servers 100 should be adapted to the specification of the SRM 110 in order to exchange, e.g., a message and data therebetween.

It is, therefore, a primary object of the present invention to provide a method and apparatus for balancing a load of a server in a VOD system.

In accordance with one aspect of the present invention, there is provided a method for balancing a load of a server in a VOD system, wherein the VOD system has a plurality of VOD servers and a multiplicity of clients, comprising the steps of:

(a) receiving a service demand message from one of the clients;

(b) retrieving VOD server addresses;

(c) calculating a remnant transmission rate of each of the VOD servers based on the retrieved VOD server addresses;

(d) selecting one of the VOD servers capable of providing a video stream corresponding to the service demand message based on the calculation result;

(e) assigning the service demand message to the selected VOD server; and

(f) send a session-setup-response message to the client who has requested the service demand message.

In accordance with another aspect of the present invention, there is provided an apparatus for balancing a load of a server in a VOD system, comprising: a network management block for managing a network communication between a plurality of VOD servers and a SRM (Session and Resource Manager); a load balancing block for balancing loads of the plurality of VOD servers; and a storage block for storing addresses of the plurality of VOD servers and the statuses thereof.

The above and other objects and features of the present invention will become apparent from the following description of preferred embodiments given in conjunction with the accompanying drawings, in which:

Fig. 1 represents a schematic block diagram of a conventional distributed VOD system;

Fig. 2 illustrates a schematic block diagram of a distributed VOD system employing a server load adjustment block in accordance with the present invention;

Fig. 3 exemplifies a detailed block diagram of the server load adjustment block shown in Fig. 2; and

Fig. 4 is a flow chart for explaining an operation of the server load adjustment block.

One of the preferred embodiments in accordance with the present invention will be described with reference to Figs. 2 to 4. In Fig. 2, there is shown a distributed VOD system in accordance with the present invention comprising a plurality of VOD severs 100, a SRM (Session and Resource Manager) 110, a multiplicity of clients 120 and a server load adjustment block 200. Each of the plurality of VOD servers functions same way in providing a service to clients. The plurality of VOD servers 100, the SRM 110 and the multiplicity of clients 120 are identical to ones shown in Fig. 1.

The SRM 110 recognizes an address of the server load

adjustment block 200 as a basic address when a service request message is generated by one of the clients 120; and relays the request message to the VOD servers 100 through the server load adjustment block 200. That is, the SRM 110 recognizes the server load adjustment block 200 as one VOD server. Thereafter, the server load adjustment block 200 selects one of the VOD servers 100 capable of providing a video stream stored therein to the client who issued the service request message, the video stream corresponding to the service request message.

Referring to Fig. 3, there is illustrated a block diagram of the server load adjustment block 200 including a network management block 310 being bidirectionally connected to the plurality of the VOD servers 100 and the multiplicity of clients 120, a message relay block 320, a load balancing block 330, a server address storage block 340 and a server status information storage block 350.

The network management block 310 manages and controls the server load adjustment block 200 by the aid of the message relay block 320 to exchange messages between the plurality of VOD servers 100 and the SRM 110, i.e., to provide the SRM 110 with a message fed from one of the plurality of VOD server 100 and vice versa.

When VOD server address information, responsive to the service request message, is fed from the server address storage block 340 through the load balancing block 330, the

network management block 310 lets the message relay block 320 relay the service request message to one of the plurality of VOD servers 100 which corresponds to the VOD server address information. The server address storage block 340 contains server addresses of the plurality of VOD servers, wherein the server addresses are used by the load balancing block 330 in directing the service request message to one of the plurality of VOD servers.

In response to the service request message, one of the plurality of the VOD servers 100 provides the network management block 310 with a video stream corresponding to the service request message. Thereafter, the network management block 310 transmits the video stream to the SRM 110.

The message relay block 320 connected to the network management block 310 performs a function that relays a communication message between one of the plurality of VOD servers 100 and the SRM 110.

The server status information storage block 350 stores status information on each of the plurality of the VOD servers 100, wherein the status information is used in balancing a load to each of the plurality of VOD servers 100. The status information includes a maximum and a current transmission rates of the video streams in each of the plurality of VOD servers; and provides the same, if required, to the load balancing block 330.

In other words, when the service request message is

inputted by one of the plurality of clients 120 via the SRM 110, the load balancing block 330, responsive to the service request message, detects a VOD server among the plurality of VOD servers 100 capable of providing the video stream corresponding to the service request message by calculating a remnant transmission rate of each of the plurality of VOD servers 100, wherein the remnant transmission rate is computed by subtracting the current transmission rate of the server from its the maximum transmission rate. Thereafter, the load balancing block 330 provides the server address storage block 340 and the network management block 310 with the detected VOD server address.

For example, if one of the plurality of the VOD servers 100 has a maximum transmission rate of 100Mbps and a current transmission rate of 60Mbps, since the remnant transmission rate of the VOD server becomes 40Mbps, the load balancing block 330 assigns the service request message to that VOD server only when the quantity of the video stream corresponding to the service request message is less than 40Mbps.

If the load balancing block 330 assigns the service request message to that VOD server, the server address storage block 340 stores VOD server address information corresponding to that VOD server fed from the load balancing block 330 as well as parameters required in a network communication by sending a Config message to the load balancing block 330,

wherein the VOD server address information indicates that VOD server among the plurality of VOD servers being capable of providing a video stream corresponding to the service request message.

Referring to Fig. 4, there is shown a flow chart for explaining an operation of the server load adjustment block 200 shown in Fig. 3. At step 402, a service request message is inputted from one of the clients 120. Next, at step 404, a server address retrieving procedure is performed in order to select a server capable of providing a video stream corresponding to the service request message. At step 406, the load balancing block 330 shown in Fig. 3 examines the remnant transmission rate of the server selected at step 404 to decide whether or not the selected server is capable of providing the requested service, wherein the remnant transmission rate of the server is obtained by subtracting a USING_BANDWIDTH of the selected server from its MAX_BANDWIDTH, the MAX_BANDWIDTH representing the maximum transmission rate of the server and the USING_BANDWIDTH being its current transmission rate. If the examination result is positive, the process proceeds to step 408; and if otherwise, it proceeds to step 410.

At step 408, the service request message is assigned to the server selected at step 404 and the process proceeds to step 414. Thereafter, at step 414, a Session-SetUp-Response message is transmitted to the SRM 110 and the process is

terminated.

The load balancing block 330, at step 410, checks the other servers in order to find a server capable of providing the video stream among the VOD servers 100. If a candidate VOD server is found, the process is returned to step 406; and if otherwise, it proceeds to step 412. At stp 412, an error code indicating that there is no server capable of providing the video stream is generated and the process is terminated.

While the present invention has been described with respect to certain preferred embodiments only, other modifications and variations may be made without departing from the scope of the present invention as set forth in the following claims.

Claims

1. A method for balancing a load of a server in a VOD system, wherein the VOD system has a plurality of VOD servers and a multiplicity of clients, comprising the steps of:

(a) receiving a service request message from one of the clients;

(b) retrieving VOD server addresses;

(c) calculating a remnant transmission rate of each of VOD servers based on the retrieved VOD server addresses;

(d) selecting one of the VOD servers capable of providing a video stream corresponding to the service request message based on the calculation result;

(e) assigning the service request message to the selected VOD server; and

(f) send a session-setup-response message to the client who has issued the service request message.

2. The method of claim 1, wherein, at the step (c), the remnant transmission rate is obtained by subtracting a current transmission rate from a maximum transmission rate of the selected VOD server.

3. The method of claim 2, wherein the maximum transmission rate is a maximum physical transmission limitation of the selected VOD server.

4. The method of claim 3, wherein the current transmission rate is a transmission rate being currently used by the selected VOD server.

5. An apparatus for balancing a load of a server in a VOD system, comprising:

means for managing network communications between a plurality of VOD servers, a SRM (Session and Resource Manager) and a multiplicity of clients;

means for balancing loads of the VOD servers; and

means for storing addresses of the VOD servers and the statuses thereof.

6. The apparatus of claim 5, wherein the balancing means calculates a remnant transmission rate of each of the VOD servers by subtracting its current transmission rate from its maximum transmission rate.

7. The apparatus of claim 6, wherein the balancing means selects one of the VOD servers based on the calculated remnant transmission rate and assigns a service request message to the selected VOD server, the service request message being issued by one of the clients.

8. The apparatus of claim 7, wherein the maximum transmission rate is a maximum physical transmission

limitation of the selected VOD server.

9. The apparatus of claim 8, wherein the current transmission rate is a transmission rate being currently used by the selected VOD server.

10. A method for balancing a load of a server substantially as herein described with reference to or as shown in Figures 1 to 4 of accompanying drawings.

11. An apparatus for balancing a load of a server constructed and arranged substantially as herein described with reference to or as shown in Figures 1 to 4 of accompanying drawings.

The Patent Office

13

INVESTOR IN PEOPLE

Application No: GB 9828598.4 **Examiner:** M J Billing
Claims searched: 1 to 4 **Date of search:** 25 May 1999

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.Q): H4K KOD3; H4R RCSS, RCST, RCT, RCX.

Int Cl (Ed.6): H04N 7/173.

Other: ONLINE - EPODOC, WPI.

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
A	EP0770964A1	(MATSUSHITA) - Figs.26-31	1
A	EP0767585A2	(IBM) - Figs.1,2A	1
A	US5737747	(EMC) - Figs.17-20, column 23 line 20 to column 25 line 12	1
A	US5630007	(MITSUBISHI) - Figs.7,12; column 8 lines 50-54	1

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.